Construction d'un gold standard pour les données agronomiques

Imène Chentli, Pierre Larmande, Konstantin Todorov

Construction d'un *gold standard* pour les données agronomiques

Imène Chentli[1,2,3], Pierre Larmande[3] et Konstantin Todorov[1,2]

[1] Université de Montpellier,
chentli.imene@hotmail.fr
[2] Laboratoire d'Informatique, Robotique et Micro-électronique de Montpellier (LIRMM),
Konstantin.Todorov@lirmm.fr
[3] Institut de Biologie Computationnelle (IBC), Montpellier,
pierre.larmande@ird.fr

Résumé : Dans le contexte des ressources agronomiques et au jour d'aujourd'hui, il n'existe ni de données de référence ni de questions adaptées au domaine pour évaluer les interfaces dotées de Systèmes à Questions-Réponses (SQR). A cet effet, nous avons construit un étalon-or (*gold standard*) constitué d'un ensemble de questions de référence formulées en langage naturel et leur correspondance en langage SPARQL. Les ressources d'Agronomic LinkedData constituent le jeu de données de référence pour la grande richesse de la base de connaissances.

Mots-clés : Données de référence, Web Sémantique, Agronomie, Interrogation de données, Systèmes de Questions-Réponses, Traitement de Langage Naturel, SPARQL.

1. Introduction

La recherche d'information est une étape laborieuse qui requiert le tri et la précision. De ce fait, la nouvelle génération de web de données liées a pour but de lever ces difficultés grâce à une représentation sémantique des connaissances stockées en format RDF[1]. Il existe à ce jour différentes bases de connaissances selon le domaine. En biologie par exemple, les principales problématiques sont de comprendre comment les experts vont pouvoir accéder à ces données et comment être sûr qu'il n'en manque pas dans les résultats proposés vu la diversité et la multiplicité des ressources. En effet, les sciences de la vie dans le web de données liées comprennent 192 millions de liens pointant vers d'autres domaines et plus de 3 milliards de triplets RDF soit environ 10% du nombre total de triplets de tous les domaines confondus. En trois années, l'ensemble de données passe de 41 à 85 ensembles (Schmachtenberg, 2014). Ces ressources structurées sont accessibles et mises à jour régulièrement pour leur majorité *via* différentes Interfaces à Langage Naturel (ILN). Ces dernières permettent aux experts de profiter de l'expressivité puissante des standards du Web Sémantique (WS) sans se soucier de leur complexité.

Ainsi, les biologistes expriment leurs besoins en LN. Puis, l'ILN les *traduit* en un langage formel. Et lorsqu'il est possible, la machine propose une réponse adaptée. A cet effet, le passage du LN au langage de machine nécessite une étape intermédiaire de *traduction* ou de *formalisation*. Malgré la maturité du WS, la difficulté réside non seulement dans la localisation des ressources mais également dans leur connexion les unes aux autres au sein des différentes bases de connaissances. L'accès aux connaissances reste une problématique non résolue dépendant de l'expertise du biologiste. Le langage SPARQL[2]; standard

[1] Ressource Description Framework est un modèle de graphe décrivant les ressources du Web sous forme de triplets (Sujet, Prédicat, Objet) et un langage de base du Web Sémantique

[2] SPARQL Protocol and RDF Query Language

recommandé par le W3C[3], est utilisé comme langage formel pour l'interrogation, l'accès aux bases de connaissances et *in fine,* la traduction du LN. Cependant, sa complexité peut restreindre son utilisation. Ainsi, la semi-automatisation de la traduction des requêtes en LN vers le SPARQL fait l'objet de recherche dans différents domaines. L'intérêt est d'éviter aux biologistes et autres utilisateurs non-spécialistes, la confrontation à la complexité du SPARQL notamment pour sa syntaxe et l'organisation des données dans les bases de connaissances.

Notre travail consiste en la construction d'un *gold standard* dédié à la communauté agronomique car jusqu'à présent, il n'existe ni de données de références ni de questions préétablies pour l'évaluation des systèmes les exploitant. Nous proposons ainsi les ressources d'Agronomic LinkedData (AgroLD) comme jeu de données de référence et un ensemble de 50 questions construites en LN et leur correspondance en langage SPARQL. Cette première étape permettra la préparation aux phases de tests et d'évaluation des ILN.

2. Travaux existants

Il existe des ILN qui emploient des Systèmes à Questions-Réponses (SQR) dépendant du degré d'expertise pour le domaine et le langage employé (LN ou SPARQL par exemple). D'autres dépendent de la liberté de son utilisation pouvant aller d'une interface proposant un formulaire ou un simple champs textuel jusqu'aux interfaces les plus complexes où l'information est visualisée dynamiquement. L'architecture d'un SQR est généralement composée de trois modules. Le premier analyse la question, le deuxième sélectionne un ensemble de résultats candidats et le dernier analyse les candidats et en extrait les réponses les plus pertinentes si elles sont présentes. Notre attention est portée sur des SQR utilisant une voire plusieurs approches de traduction du LN en requêtes SPARQL pour interroger des bases de connaissances structurées. Notre champs d'études considère tous les types de requêtes (mots clés, fragments ou phrase intégrales) et il est élargi à tous les domaines d'application car au jour d'aujourd'hui, il existe très peu de SQR dédiés au domaine biologique (Neves & Leser, 2015). Des travaux de traduction du LN vers des requêtes sémantiques ont mis en évidence l'emploi de différentes approches. Nous en énumèrerons les principales en nous limitant à un exemple d'ILN par approche :

1. **Approche basée sur des requêtes en LN contrôlé** - Ginseng (Bernstein et *al*., 2005) peut contrôler les requêtes selon l'ontologie fournie à l'outil. Ginseng utilise des règles statiques définissant la structure générale de la requête et dynamiques issues des étiquettes de la base de connaissances. Cependant, l'expert est limité dans la manière de rédiger ses requêtes. Lorsque le terme n'apparaît pas dans la liste déroulante cela signifie que le système considère que sa grammaire est incorrecte et ne l'accepte pas.
2. **Approche basée sur l'alignement de structures sémantiques aux ontologies -** Querix (Kaufmann et *al.,* 2006) indépendant du domaine, analyse la structure syntaxique de la requête pour trouver de meilleures correspondances avec la base de connaissances. Le système interagit avec l'utilisateur à l'aide d'une fenêtre de dialogue pour lui proposer différents triplets. Lorsque l'utilisateur les sélectionne, une requête en SPARQL est générée et exécutée. Cependant, le système ne résout pas les ambiguïtés. Les requêtes sont en Anglais et doivent impérativement commencer par une question à Wh.
3. **Approche basée sur une grammaire formelle -** ORAKEL (Cimiano, et *al.*, 2007) utilise deux types de lexiques lors du processus d'analyse. Le premier est général et indépendant du domaine pour la construction de la requête. Le second correspond au domaine. ORAKEL a donc besoin de connaissances ontologiques pour interpréter sémantiquement des termes ou pour résoudre des ambiguïtés. Si la source ontologique est incomplète, les résultats le seront aussi lors de son évaluation.

---

[3] World Wide Web Consortium : http://www.w3.org

4. **Approche basée sur des patrons -** LODQA (Kim & Cohen, 2013) analyse les requêtes d'un point de vue linguistique, aligne les phrases nominales aux différents termes ontologiques et propose des résultats en SPARQL. L'avantage repose sur sa modulabilité pour analyser la requête. Par ailleurs, l'interface reste en cours de développement et limitée à une ontologie.
5. **Approche basée sur l'auto-génération de requêtes SPARQL -** Sparklis (Ferré, 2014) est indépendant du domaine. L'outil permet d'exploiter différents types de bases de connaissances *via* un point d'accès SPARQL en guidant l'utilisateur pas à pas à construire les questions, les requêtes en SPARQL et obtenir les réponses de façon interactive. Le système donne également des suggestions pour raffiner la sélection.

Plusieurs campagnes d'évaluation de SQR sont menées régulièrement telles que QALD[4] et BioNLP[5] pour le domaine biomédical. Par ailleurs, les challenges de SQR sont inexistants pour les ressources agronomiques. Chaque domaine possédant sa propre terminologie, il serait alors difficile de tester un SQR avec un même jeu de requêtes. Le portage? de SQR vers le domaine agronomique nécessite donc la mise en place et de l'analyse d'un nouveau jeu de requêtes.

3. Démarche

Notre travail consiste à mettre en place un *gold standard* à la disposition de la communauté des plantes pour faciliter l'évaluation des SQR en respectant les besoins exprimés par les experts et leur capacité à faire face aux grands volumes hétérogènes de données. L'objectif final consiste à ce que le modèle à tester retourne, pour une question en LN et une source de données RDF, une liste d'entrées répondant aux questions posées. Afin de mettre en place le *gold standard*, nous avons tenu compte des éléments suivants :

a. **L'ensemble de données.** Ce sont des données interconnectées en graphes RDF. Nous exploiterons la base de connaissances du projet AgroLD (Venkatesan et al., 2015). En effet, elle est accessible à toute la communauté et représente une grande diversité de données agronomiques.
b. **L'infrastructure.** Dans notre cas, elle correspond au point d'accès SPARQL d'AgroLD directement accessible en ligne[6].
c. **Un *gold standard*** permettant d'effectuer des tests d'entraînement à l'aide de l'ensemble de données. Pour ce faire, nous avons construit manuellement 50 questions en LN et 50 requêtes en SPARQL correspondantes. Ces dernières sont soumises au point d'accès SPARQL d'AgroLD pour extraire les réponses correspondantes. Une fois validé, ce *gold standard* servira *in fine* de référence.
d. **Un modèle** correspond à un voire une liste de SQR à évaluer selon une procédure et des mesures d'évaluation.

4. Conclusion

Nous avons vu que pour différentes approches, il existe un grand nombre d'interfaces avec un SQR proposant la formalisation du LN dont la majorité dépend du domaine. En effet, il est très difficile de trouver un SQR générique. Les principales raisons sont le jeu de données et la richesse des liens qu'il peut avoir avec différentes bases de connaissances.

---

[4] http://www.sc.cit-ec.uni-bielefeld.de/qald

[5] http://2016.bionlp-st.org

[6] Point d'accès SPARQL des données d'AgroLD : http://volvestre.cirad.fr:8080/aldp/

L'objectif de notre travail est la mise en place d'un *gold standard* pour la communauté agronomique afin de booster le développement de solutions et de nouvelles approches adaptées à ce domaine et de lancer de nouvelles campagnes d'évaluations de SQR.

Références

BERNSTEIN A., KAUFMANN E., KISER C., KIEFER C., (2006). Ginseng: A Guided Input Natural Language Search Engine for Querying Ontologies, 2006 Jena User Conference, Bristol, UK, May 2006.

CIMIANO P., HAASE P. & HEIZMANN J. (2007). Porting natural language interfaces between domains : An experimental user study with the orakel system. In Proceedings of the 12th international conference on intelligent user interfaces (pp. 180-189). ACM.

FERRÉ S. (2014). SPARKLIS : a SPARQL endpoint explorer for expressive question answering. In ISWC posters & demonstrations track (Vol. 1272, pp. 45-48).

KAUFMANN E., BERNSTEIN A. & ZUMSTEIN R. (2006). Querix : A natural language interface to query ontologies based on clarification dialogs. Springer.

KIM J.D. & COHEN K. (2013). Natural language query processing for sparql generation : A prototype system for SNOMEDCT. In Proceedings of the BioLINK SIG (pp. 32-38).

NEVES M. & LESER U. (2015). Question answering for Biology. Methods, 74, 36-46.

SCHMACHTENBERG M., BIZER C., PAULHEIM H. (2014) Adoption of the Linked Data Best Practices in Different Topical Domains. 13th International Semantic Web Conference (ISWC2014) - RDB Track (2014)

VENKATESAN A., LARMANDE P., JONQUET C., RUIZ M. & VALDURIEZ P. (2015). Facilitating efficient knowledge management and discovery in the Agronomic Sciences. In Proceedings of the 8th Semantic Web Applications and Tools for Life Sciences International Conference, Cambridge UK, December 7-10, 2015 (pp. 205-207).